PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G10L 3/02 — A1

(11) International Publication Number: WO 99/66494

(43) International Publication Date: 23 December 1999 (23.12.99)

(21) International Application Number: PCT/US99/12804

(22) International Filing Date: 16 June 1999 (16.06.99)

(30) Priority Data:
09/099,952 19 June 1998 (19.06.98) US

(71) Applicant: COMSAT CORPORATION [US/US]; 6560 Rock Spring Drive, Bethesda, MD 20817 (US).

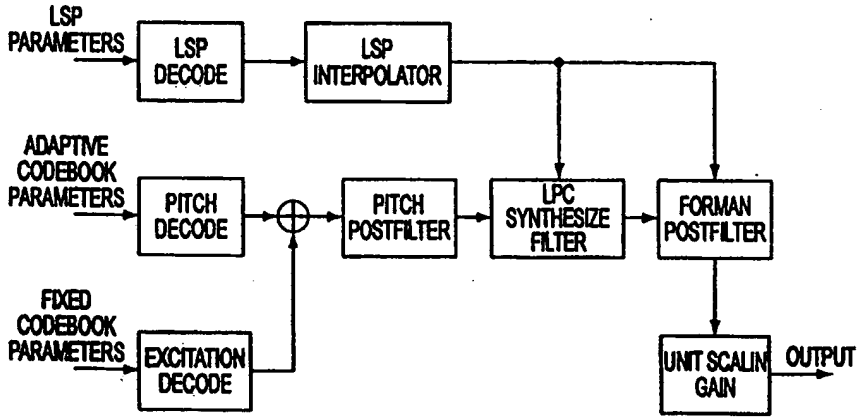
(72) Inventors: HO, Grant, Ian; 84 Noth Hills Terrace, Don Mills, Ontario M3C 1M6 (CA). BARANIECKI, Marion; 4781 Farndon Court, Fairfax, VA 22032 (US). YELDENER, Suat; 19606 Crystal Rock Drive #14, Germantown, MD 20874 (US).

(74) Agents: CUSHING, David, J. et al.; Sughrue, Mion, Zinn, MacPeak & Seas, PLLC, Suite 800, 2100 Pennsylvania Avenue, N.W., Washington, DC 20037-3202 (US).

(81) Designated States: AU, CA, IN, European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: IMPROVED LOST FRAME RECOVERY TECHNIQUES FOR PARAMETRIC, LPC-BASED SPEECH CODING SYSTEMS



(57) Abstract

A lost frame recovery technique for LPC-based systems employs interpolation of parameters from previous and subsequent good frames, selective attenuation of frame energy when the energy of a subframe exceeds a threshold, and energy tapering in the presence of multiple successive lost frames.

IMPROVED LOST FRAME RECOVERY TECHNIQUES FOR PARAMETRIC, LPC-BASED SPEECH CODING SYSTEMS

Background of the Invention

The transmission of compressed speech over packet-switching and mobile communications networks involves two major systems. The source speech system encodes the speech signal on a frame by frame basis, packetizes the compressed speech into bytes of information, or packets, and sends these packets over the network. Upon reaching the destination speech system, the bytes of information are unpacketized into frames and decoded. The G.723.1 dual rate speech coder, described in *ITU-T Recommendation G.723.1*, "Dual Rate Speech Coder for Multimedia Communications Transmitting at 5.3 and 6.3 kbit/s," March 1996 (hereafter "Reference 1", and incorporated herein by reference) was ratified by the ITU-T in 1996 and has since been used to add voice over various packet-switching as well as mobile communications networks. With a mean opinion score of 3.98 out of 5.0 (see, Thryft, A. R., "Voice over IP Looms for Intranets in '98," *Electronic Engineering Times*, August, 1997, Issue: 967, pp. 79, 102, hereafter "Reference 2", and incorporated herein by reference), the near toll quality of the G.723.1 standard is ideal for real-time multimedia applications over private and local area networks (LANs) where packet loss is minimal. However, over wide area networks (WANs), global area networks (GANs), and mobile communications networks, congestion can be severe, and packet loss may result in heavily degraded speech if left untreated. It is therefore necessary, to develop techniques to reconstruct lost speech frames at the receiver in order to minimize distortion and maintain output intelligibility.

The following discussion of the G.273.1 dual rate coder and its error concealment will assist in a full understanding of the invention.

The G.723.1 dual rate speech coder encodes 16-bit linear pulse-code modulated (PCM) speech, sampled at a rate of 8 KHz, using linear predictive analysis-by-synthesis coding. The excitation for the high rate coder is Multipulse Maximum Likelihood Quantization (MP-MLQ) while the excitation for the low rate coder is Algebraic-Code-Excited Linear-Prediction (ACELP). The encoder operates on a 30

the average of the gains for subframes 2 and 3 of the previous frame. Otherwise, for the voiced case, the previous frame is attenuated by 2.5 dB and regenerated with a periodic excitation having a period equal to the estimated pitch lag. If packet losses continue for the next two frames, the regenerated excitation is attenuated by an additional 2.5 dB for each frame, but after three interpolated frames, the output is completely muted, as described in Reference 1.

The G.723.1 error concealment strategy was tested by sending various speech segments over a network with packet loss levels of 1%, 3%, 6%, 10%, and 15%. Single as well as multiple packet losses were simulated for each level. Through a series of informal listening tests, it was shown that although the overall output quality was very good for lower levels of packet loss, a number of problems persisted at all levels and became increasingly severe as packet loss increased.

First, parts of the output segment sounded unnatural and contained many annoying, metallic-sounding artifacts. The unnatural sounding quality of the output can be attributed to LSP vector recovery based on a fixed predictor as previously described. Since the missing frame's LSP vector is recovered by applying a fixed predictor to the previous frame's LSP vector, the spectral changes between the previous and reconstructed frames are not smooth. As a result of the failure to generate smooth spectral changes across missing frames, unnatural sounding output quality occurs, which increases unintelligibility during high levels of packet loss. In addition, many high-frequency, metallic-sounding artifacts were heard in the output. These metallic-sounding artifacts primarily occur in unvoiced regions of the output, and are caused by incorrect voicing estimation of the previous frame during excitation recovery. In other words, since a missing, unvoiced frame may incorrectly be classified as voiced, then transition into the missing frame will generate a high-frequency glitch, or metallic-sounding artifact, by applying the estimated pitch lag computed for the previous frame. As packet loss increases, this problem becomes even more severe, as incorrect voicing estimation generates increased distortion.

Another problem using G.723.1 error concealment was the presence of high-energy spikes in the output. These high-energy spikes, which are especially

by applying the second part of the linear interpolation technique, almost all unwanted metallic-sounding artifacts are effectively masked away.

To eliminate the effects of high-energy spikes, a selective energy attenuation technique was developed. This technique checks the signal energy for every synthesized subframe against a threshold value, and attenuates all signal energies for the entire frame to an acceptable level if the threshold is exceeded. Combined with linear interpolation, this selective energy attenuation technique effectively eliminates all instances of high-energy spikes from the output.

Finally, an energy tapering technique was designed to eliminate the effects of "choppy" speech. Whenever multiple packets are lost in excess of one frame, this technique simply repeats the previous good frame for every missing frame by gradually decreasing the repeated frame's signal energy. By employing this technique, the energy of the output signal is gradually smoothed or tapered over multiple packet losses, thus eliminating any patches of silence or a "choppy" speech effect evident in G.723.1 error concealment. Another advantage of energy tapering is the relatively small amount of computation time required for reconstructing lost packets. Compared to G.723.1 error concealment, since this technique only involves gradual attenuation of the signal energies for repeated frames, as opposed to performing G.723.1 fixed LSP prediction and excitation recovery, the total algorithmic delay is considerably less.

Brief Description of the Drawing

The invention will be more clearly understood from the following description in conjunction with the accompanying drawing, wherein:

Fig. 1 is a block diagram showing G.723.1 decoder operation;

Fig. 2 is a block diagram illustrating the use of Future, Ready and Copy buffers in the interpolation technique according to the present invention;

Figs. 3a-3c are waveforms illustrating the elimination of high energy spikes by the error concealment technique of the present invention; and

current frame, is a good or missing frame that is currently being processed by the decoder, and is stored in the Ready Buffer.

future frame, is a good or missing frame immediately following the current frame, and is stored in the Future Buffer.

Linear interpolation is a multi-step procedure that operates as follows:

1. The Ready Buffer stores the current good frame to be processed while the Future Buffer stores the future frame of the encoded speech sequence. A copy of the current frame's speech model parameters is made and stored in the Copy Buffer.

2. The status of the future frame, either good or missing, is determined. If the future frame is good, no linear interpolation is necessary; and the linear interpolation flag is reset to 0. If the future frame is missing, linear interpolation might be necessary; and the linear interpolation flag is temporarily set to 1. (In a real-time system, a missing frame is detected by either a receiver timeout or Cyclical Redundancy Check (CRC) failure. These missing frame detection algorithms however, are not part of the invention, but must be recognized and incorporated at the decoder for proper operation of any packet reconstruction strategy.)

3. The current frame is decoded and synthesized. A copy of the current frame's LPC synthesis filter and pitch postfiltered excitation are made.

4. The future frame, originally in the Future Buffer, becomes the current frame and is stored in the Ready Buffer. The next frame in the encoded speech sequence arrives as the future frame in the Future Buffer.

5. The value of the linear interpolation flag is checked. If the flag is set to 0, the process jumps back to step (1). If the flag is set to 1, the process jumps to step (6).

6. The status of the future frame is determined. If the future frame is good, linear interpolation is applied; the linear interpolation flag remains set to

13. The future frame, originally in the Future Buffer, becomes the current frame and is stored in the Ready Buffer. The next frame in the encoded speech sequence arrives as the future frame in the Future Buffer. The process then returns to step (1).

There are at least two important advantages of linear interpolation over G.723.1 error concealment. The first advantage occurs in step (7), during LSP recovery. In Step (7), since linear interpolation determines the missing frame's LSP parameters based on the previous and future frames, this provides a better estimate for the missing frame's LSP parameters, thereby enabling smoother spectral changes across the missing frame, than if fixed LSP prediction were simply used, as in G.723.1 error concealment. As a result, more natural sounding, intelligible speech is generated, thereby increasing comfortability for the listener.

The second advantage of linear interpolation occurs in steps (8) to (11), during excitation recovery. First, in step (8), since linear interpolation generates the missing frame's gain parameters by averaging the fixed codebook gains between the previous and future frames, it provides a better estimate for the missing frame's gain, as opposed to the technique described in G.723.1 error concealment. This interpolated gain, which is then applied for unvoiced frames in step (10), thereby generates smoother, more comfortable sounding gain transitions across frame erasures. Secondly, in step (11), voicing classification is based on the both the predictor gain and estimated pitch lag, as opposed to the predictor gain alone, as in G.723.1 error concealment. That is, frames whose predictor gain is greater than 0.58 dB are also compared against a threshold pitch lag, P_{thresh}. Since unvoiced frames are primarily composed of high-frequency spectra, those frames that have low estimated pitch lags, and hence, high estimated pitch frequencies, thereby have a higher probability of being unvoiced. Thus, frames whose estimated pitch lags fall below P_{thresh} are declared unvoiced and those whose estimated pitch lags exceed P_{thresh}, are declared voiced. In sum, by selectively determining a frame's voicing classification based on both the predictor gain and estimated pitch lag, the technique of this invention effectively masks away all occurrences of high-frequency, metallic-sounding artifacts

copy of the current frame's speech model parameters is made and stored in the Copy Buffer.

2. The status of the future frame, either good or missing, is determined. If the future frame is good, no linear interpolation is necessary; the linear interpolation is reset to 0. If the future frame is missing, linear interpolation might be necessary; the linear interpolation flag is temporarily set to 1.

3. The current frame is decoded and synthesized. A copy of the current frame's LPC synthesis filter and pitch postfiltered excitation is made.

4. The future frame, originally in the Future Buffer, becomes the current frame and is stored in the Ready Buffer. The next frame in the encoded speech sequence arrives as the future frame in the Future Buffer.

5. The value of the linear interpolation flag is checked. If the flag is set to 0, the process jumps back to step (1). If the flag is set to 1, the process jumps to step (6).

6. The status of the future frame is determined. If the future frame is good, linear interpolation is applied as described in subsection 3.1. If the future frame is missing, energy tapering is applied; the energy tapering flag is set to 1, the linear interpolation flag is reset to 0, and the process jumps to step (7).

7. The copy of the previous frame's pitch postfiltered excitation, from step (3), is attenuated by (0.5 × value of energy tapering flag) dB.

8. The copy of the previous frame's LPC synthesis filter, from step (3), is used to synthesize the current frame using the attenuated excitation in step (7).

9. The future frame, originally in the Future Buffer, becomes the current frame and is stored in the Ready Buffer. The next frame in the encoded speech sequence arrives as the future frame in the Future Buffer.

10. The current frame is synthesized using steps (7) to (9), then jumps to step (11).

more natural sounding speech and effective masking away of all metallic-sounding artifacts were achieved due to smoother spectral transitions across missing frames based on linear interpolation and improved voicing classification. Secondly, all high-energy spikes were eliminated due to selective energy attenuation and linear interpolation. Finally, all instances of "choppy" speech were eliminated due to energy tapering. It is important to realize that as network congestion levels increase, the amount of packet loss also increases. Thus, in order to maintain real-time speech intelligibility, it is essential to develop techniques to successfully conceal frame erasures while minimizing the amount of degradation at the output. The strategies developed by the authors represent techniques which provide improved output speech quality, are most robust in the presence of frame erasures compared to the techniques described in Reference 1, and can be easily applied with any parametric, LPC-based speech coder over any packet-switching or mobile communications network.

It will be appreciated that various changes and modifications may be made to the specific embodiments described above without departing from the spirit and scope of the invention as defined in the appended claims.

5. A method according to claim 1, wherein on loss of multiple successive frames, said method comprises the step of repeating the encoded signals for a frame immediately preceding said multiple successive frames while gradually reducing the signal energy for each recovered frame.

6. A method according to claim 2, wherein said encoded signals include said LSP parameters, fixed codebook gains and further excitation signals, said method comprising interpolating said fixed codebook gain of said lost frame from the fixed codebook gains of said first and second frames, and adopting said further excitation signals from said first frame as the further excitation signals of said lost frame.

7. A method of recovering a lost frame in a system of the type wherein information is transmitted as successive frames of encoded signals and the information is reconstructed from said encoded signals at a receiver, said method comprising:

calculating an estimated pitch value and predictor gain for a first frame prior to said lost frame; and

classifying said lost frame as voiced or unvoiced in accordance with said predictor gain and estimated pitch value from said first frame.

8. A method of recovering a lost frame in a system of the type wherein information is transmitted as successive frames of encoded signals, each frame including plural subframes, and the information is reconstructed from said encoded signals at a receiver, said method comprising:

comparing a signal energy for each subframe of a particular frame against a threshold; and

attenuating signal energies for all subframes in said particular frame if the signal energy in any subframe exceeds said threshold.

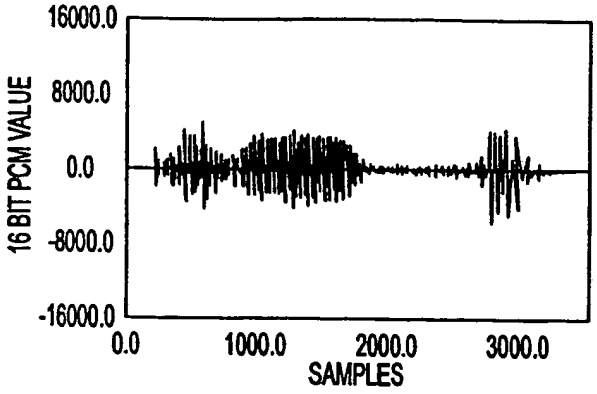


FIG. 3A

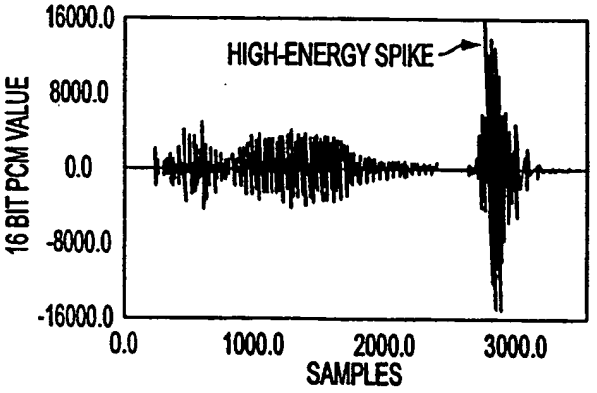


FIG. 3B

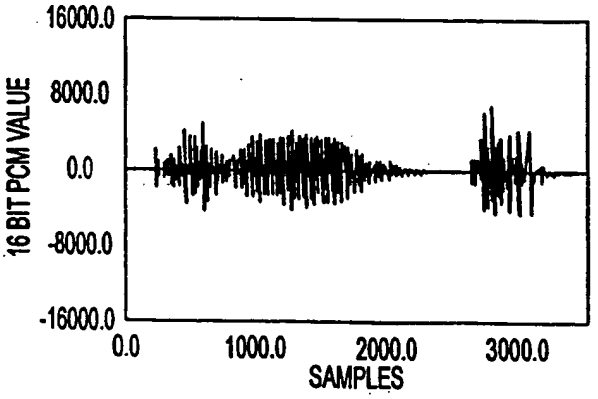


FIG. 3C

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US99/12804

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : G10L 3/02
US CL :704/223,219,201

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 704/223,219,201

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

IEL

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5,699,485 A (SHOHAM) 16 December 1997, col. 3 lines 50-60; col. 5 lines 10-30; col. 6 lines 1-16, lines 29-42; col. 7 line 60 - col. 8 line 9; col. 8 lines 30-35; col. 18 lines 14-24	1-8
X	US 5,732,389 A (KROON et al) 24 March 1998, col. 7-8	1-8
X	US 4,975,956 A (LIU et al) 04 December 1990, Abstract, Fig. 1	1-2

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"B"	earlier document published on or after the international filing date	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"&"	document member of the same patent family
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search	Date of mailing of the international search report
12 AUGUST 1999	18 OCT 1999
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer DAVID R. HUDSPETH Telephone No. (703) 308-4825

Form PCT/ISA/210 (second sheet)(July 1992)*